CLUSTER ANALYSIS OF PREVIOUSLY DESCRIBED COMMUNITIES FROM THE LUDLOW OF THE WELSH BORDERLAND

by PIERRE J. LESPÉRANCE

ABSTRACT. Previously described communities ('associations') from the Ludlow of the Welsh Borderland were subjected to cluster analysis on a PC microcomputer. Miscellaneous absence–presence data sets derived from the published information were analysed using different combinations of clustering algorithms, distance and similarity measurements, with and without Jaccard's coefficient of association. Variations in order of data entry produced major differences (chaotic behaviour) in clustering using unmodified data. With unmodified data, only the single linkage method showed no clustering differences with the three distance and similarity measurements applied to the whole of the data. The raw data, modified with Jaccard's coefficient, showed major improvement in clustering upon variation of order of data entry. Such improvement in non-chaotic behaviour is interpreted as the result of the lesser, finite raw data consisting of zeros and ones, as compared with the more infinite numbers generated using a coefficient of association. Nearly equally good results are obtained, however, when the unmodified data are analysed using the *cosine θ* measurement. The UPGMA, single and complete linkage methods, with *cosine θ*, are recommended as quickly and routinely applicable to unmodified data for, at least, first approximations in community analysis. Dendrograms so generated should nonetheless be submitted to variation in order of data entry to test for chaotic behaviour.

WITH the advent of microcomputers and powerful programs adapted to this specific hardware, treatment of data previously available only on mainframe computers is now possible within each scientist's office or home. The purpose of this contribution is to explore the use of cluster analysis on a microcomputer well within its practical limits, but in a situation typical of most palaeoecological studies.

Cluster analysis is a non-parametric statistical technique, that is, it is not based on the familiar Gaussian, or probability, bell-shaped curve. It classifies data and produces, if desired, dendrograms (i.e. tree-shaped diagrams), which are visually instructive and, hopefully, of easy interpretation of the classified data. The practical limits of the microcomputers previously alluded to, refer to the available memory. More specifically, the internal memory to hold data, or matrices, places an upper limit to these data in better equipped standard IBM-type personal microcomputers (PC) in the neighbourhood of 4·6 Mb (mega bytes) (of which 4 Mb is expanded memory), at least with the software here employed. About three-quarters of this memory is needed, in a study like this one, if a coefficient of association is calculated on a spreadsheet contained in a single file (note 1, appendix gives more details). Data storage is also a problem, but hard disks of 30 Mb are readily available, and those ten times that amount equally available, but expensive.

The description of the Ludlow communities of part of the Welsh Borderland, a portion of the world standard of the upper Silurian, was first outlined by Calef and Hancock (1974), criticized by Lawson (1975), reinvestigated by Watkins (1979) and partly extended to a regional scale by Cherns (1988). Watkins (1979) is a comprehensive study of many aspects of synecology of the whole of the Ludlow in the Welsh Borderland. Although more restricted in area than Cherns (1988), published faunal lists, community tabulations, and stratigraphical coverage are more extensive in Watkins (1979) and better suited to the aims of this study. Although he named the resultant communities 'associations', this term was used interchangeably with 'communities' (Watkins 1979, p. 210), which the writer prefers (but then, discussion persists as to whether these 'associations' are

[Palaeontology, Vol. 33, Part 1, 1990, pp. 209–224.]

communities: Cherns 1988, p. 488). Other aspects of Ludlow palaeoecology are treated in Watkins (1978), Watkins and Aithie (1980), Hewitt and Watkins (1980), and Mikulic and Watkins (1981), but these contributions do not present data as complete as in Watkins (1979), nor do they specifically modify the 1979 tabulations. Consequently, Watkins (1979) will be used exclusively here.

Stratigraphic nomenclature of the Welsh Borderland dates back, of course, to R. I. Murchison, but modern usage rests on Holland *et al.* (1963), subsequently very slightly modified by Holland *et al.* (1980), Holland (1980) and Antia (1980). As Watkins's (1979) distributional data rest on Holland *et al.* (1963), their stratigraphic nomenclature is followed.

This contribution proposes to use Watkin's (1979) data as a test case for isolating specific methods and procedures of cluster analysis. As these communities were described using 'classical' methods (i.e. communities are recurrent associations defined on the basis of abundance of specific taxa, with consideration commonly given to the spatial continuity of the associations and the absence of specific taxa), comparison of clustering efficiency and correctness can be assessed and judgments expressed. Clustering algorithms are numerous and their respective merits and disadvantages under specific circumstances are not obvious to the applied researcher (some theoretical aspects are covered in Milligan (1980) and Milligan and Isaac (1980)). Furthermore, many distinct coefficients of association between samples have been suggested (binary (absence–presence) ones are surveyed in Cheetham and Hazel 1979; a comprehensive survey is given in Legendre and Legendre 1983), but their respective advantages and disadvantages, again in specific circumstances, are equally far from obvious. In fact, one of the aims of this investigation was to question the necessity of the use of coefficients of association, following results and methodological uncertainties inherent to these results, previously obtained by Lespérance and Sheehan (1988). In view of these uncertainties, a pragmatic approach was best indicated; this is detailed in the following pages.

METHODS

Hardware

Calculations were performed on an IBM-compatible PC, equipped with an 8088 chip and mathematical coprocessor (8087). Mainboard memory was 640 Kb, with an expanded memory card of 1·5 Mb (only used by the Symphony software). The PC had a 30 Mb hard disk, with a tape backup of 60 Mb. An EGA card (Extended Graphics Adapter) or better, with its consequent monitor, are a requisite to produce the text-figures as presented. Mainboard memory was always sufficient, and matrices 2·4 times the size of the data here treated, requiring 188 K of RAM, have been analysed without memory shortage, although in this last case processing time increases dramatically to about $\frac{1}{2}$ hour.

Software

Statistical calculations were done using the SPSS/PC+statistical package (version 2·0). Cluster analysis is available in the advanced statistics package; the optional data entry package facilitated the entry of the unmodified data from Watkins (1979), subsequent modification in a spreadsheet, and/or direct entry in the statistical programs. Jaccard's coefficient of association (note 1, appendix) was calculated on a spreadsheet; Symphony, version 1·2, was used (Lotus 1-2-3 is equivalent). A file compression utility (SQZ! plus) was invaluable to manage the matrices generated to calculate Jaccard's coefficient.

Data sets

The justification for the analysis of different data sets, all derived from Watkins (1979), will be presented in subsequent sections. Only a brief summary of the relations between the major data sets is presented here.

Tables 15 to 20 of Watkins (1979) give detailed faunal lists of the six communities, and intermediates or variants, recognized in the Ludlow of the Welsh Borderland. The communities are not detailed in the same fashion; as a first step, the 48 different faunal lists were included in one data set. Only taxa identified to the specific or generic level were retained for analysis. A few pelagic taxa were singled out by Watkins (1979); these were included, if only because they provide some sort of information on the physico-chemical conditions in the water column above the level-bottom communities, if indeed all the taxa so identified were pelagic (as, for instance, the case of the brachiopod *Aegiria grayi*: Cherns 1988, p. 486). A total of 112 taxa was consequently retained and, unfortunately, the *Atrypa reticularis*-coral community does not contain a generically determined coral with this procedure. The order of data entry was as presented successively in Watkins's (1979) tables 15 to 20, except that three of the six cumulative faunas, of medium to high diversity from table 15, were arbitrarily entered as the last three faunas. The first data set is referred to as data set 48A. Data set 48B is identical to 48A except that five faunal lists, chosen at random (with the Symphony function of the same name) were removed and reinserted as the last five faunal lists.

Data set 46A was derived from 48A by the deletion of two cumulative faunas from Watkins' (1979) table 15: the *Go2* (*Glassia obovata*) fauna in the mudstone facies is detailed in table 17 and the *Ml* (*Mesopholidostrophia laevigata*) fauna in table 18, and hence repetitive in the data set. Data set 46B was derived from 48B, with the deletion of the same two cumulative faunas.

Table 16 of Watkins (1979) lists six cumulative faunas of two taphonomic categories: disturbed neighbourhood assemblages, and transported assemblages, as recognized in three different communities. The faunal content of these three communities, from specific localities and samples, is presented in subsequent tables. Hence, these six faunal lists are also repetitive in the data sets, and they were deleted to produce data set 40. Data set 40A was derived from 46A, and 40B from 46B. Data set 40C is data set 40B, with five faunal lists chosen at random, deleted, and reinserted as the last five entries. As selected faunal lists were deleted, so were the taxa occurring only in the deleted faunal lists.

The raw, unmodified data were coded 0 and 1 (absent and present respectively) and used as such directly in the clustering. These data were subsequently used for Jaccard's coefficient, and the same order of data entry retained (e.g. data set 46A, modified or not by Jaccard's coefficient, has the same sequential order of entry). Justification of the use of absence–presence data is found in most discussions of cluster analysis, and need not be repeated here. The writer believes it is particularly appropriate to regional palaeoecological syntheses, to nullify local effects of species abundances.

THE COMMUNITIES

Watkins (1979) investigated the Ludlow of part of the Welsh Borderland, but excluded the basal part (the Lower Elton Beds) and the overlying brackish water Ludlow Bone Bed. Within this sequence, he recognized six communities which are, in ascending order, the *Glassia obovata* (*Go*) community, a succeeding transitional (*tr*) fauna with the one above, the *Mesopholidostrophia laevigata* (*Ml*) community, a lower phase of the *Sphaerirhynchia wilsoni* (*lwSw*) community, an *Atrypa reticularis*-coral (*AC*) community, the preceeding two communities locally absent below the upper phase of the *Sphaerirhynchia wilsoni* (*upSw*) community, the *Shaleria ornatella* (*So*) community and, uppermost, the *Protochonetes ludloviensis* (*Pl*) community (table 1). These communities were defined using 'classical' methods and, more specifically, by a graphical method known as the transect method (Watkins 1979, p. 208) on a spatial basis.

Appendix 3 of Watkins (1979, tables 15–20, pp. 262–274) presents a formidable amount of distributional data, comprising but a small part of his unpublished data in the Library of the British Museum (Natural History). Visual examination of these data does not obviously reveal the community relationships, and hence its intended use as a test case for cluster analyses.

Tables 15 and 16 of Watkins (1979) give 12 cumulative faunas, while tables 17–20 give examples (from specific localities and collections) from five communities, the *Go* fauna in the laminated shale facies, the *lpSw* and the *AC* faunas being known only by cumulative faunas. Diversity (*d*) and average diversity ($\bar{d}$) within the miscellaneous faunal lists vary widely. Cumulative faunas have $d = 22$ to 75. Average diversity in tables 17–20 of Watkins (1979) decreases to about half its value

TABLE 1. Stratigraphic relations of the benthic communities discussed within the Ludlow (modified from Watkins 1979)

WHITCLIFFE	*Pl*	*Protochonetes ludloviensis*
LEINTWARDINE	*So*	*Shaleria ornatella*
	upSw	upper phase *Sphaerirhynchia wilsoni*
BRINGEWOOD	*AC*	*Atrypa reticularis*-coral
	lpSw	lower phase *Sphaerirhynchia wilsoni*
	Ml	*Mesopholidostrophia laevigata*
ELTON	*tr*	transitional fauna
	Go	*Glassia obovata*

from the lower to the upper part of the Ludlow: from $\bar{d} = 18{\cdot}9$ in the *Go* fauna, to $\bar{d} = 15{\cdot}7$ in the *Ml* and *Sw* faunas, to $\bar{d} = 8{\cdot}8$ in the *Sw* fauna and $\bar{d} = 7{\cdot}3$ in the *Pl* fauna.

Distributional data are seldom available for sequences of communities, and Watkins's (1979) readily available publication mitigates against using his more complete unpublished data. Admittedly, these data are almost a worst case situation with such widely varying diversities (the *Pl*2C2 locality has but four taxa!) and unequal tabulation of cumulative faunas and individual localities. Nonetheless, results obtained were encouraging (see below).

CLUSTERING BEHAVIOUR: CHAOTIC?

Reproducibility is assuredly a prime requisite of any analytical method. It is thus particularly disturbing that, following F. Vogel, Bayer (1985, p. 98) has shown, at least geometrically, and with specific data sets, that cluster analysis is subject to chaotic processes (i.e. stochastic, random, aleatory processes, and hence the results are unreproducible) during the formation of clusters. This chaotic behaviour depends, to a great extent, on the sequence of input of the data. This serious defect of cluster analysis needs clarification before any palaeoecological application can confidently be pursued.

Data sets 48, 46, and 40 were used as a means of judging this suspected chaotic behaviour. Each set was used in its unmodified form, and its modified form using Jaccard's coefficient of association, and subjected to various clustering algorithms and distance, or similarity measurements.

Cluster techniques

Cluster analysis is described to varying extents by the following authors, amongst others, to which the reader may refer for fuller treatment than presented here: Sneath and Sokal (1973), Anderberg (1973), Everitt (1980), Legendre and Legendre (1983), Romesburg (1984), and Jones (1988). Q-mode analysis (between samples) and hierarchical clustering are used exclusively in this contribution.

Clustering algorithms. Seven clustering algorithms (methods for combining clusters) are available in the SPSS/PC+package. These are: (1) average linkage within groups method (WPGMA: weighted pair-group method using arithmetic averages), (2) average linkage between groups method (UPGMA: unweighted pair-group method using arithmetic averages), (3) single linkage method, (4) complete linkage method, (5) centroid method (UPGMC: unweighted pair-group centroid method), (6) median method (WPGMC: weighted pair-group centroid method), and (7) Ward's method. Much mistrust has been expressed relative to the single linkage method (discussed in Milligan 1980; rejected by Legendre *et al.* 1985, p. 275 in succession studies), although it has a natural logic, while UPGMA is generally considered to be the best method (and used exclusively, for instance, by Baarli 1987). Ward's method is popular (and used exclusively, for instance, by Lespérance and Sheehan 1988); it produces, probably, the most visually appealing (and interpretable?) dendrograms (see also discussion by Romesburg 1984, pp. 134–135).

Distance and angular measurements. Before combining clusters, an assessment of the distance between the items to be clustered is calculated. Either euclidian distances, or squared euclidian distances, can be used in the clustered hyperspace. Euclidian distances obey the familiar pythagorean relations of the hypotenuse in a triangle, while squared euclidian distances do not. Squared euclidian distance is consequently referred to as a pseudometric or a semimetric measure (Sneath and Sokal 1973, p. 121; Legendre and Legendre 1983, p. 194). Ward's method (Ward 1963) explicitly requires squared euclidian distances, while this same measure is preferable with the centroid and median methods (discussion in Sneath and Sokal 1973, p. 235). Reversals in cluster formation occur when euclidian distances are used with the centroid and median methods (Sneath and Sokal 1973, p. 235), but also occur with squared euclidian distances and the centroid method (Boyce 1969, p. 15). These reversals in clustering values (the agglomeration schedule printouts are available with the SPSS/PC+package) occur in both the raw and modified data, and have been observed with the three distance and similarity measures used with the centroid and median methods. Additional complications arise with the use of euclidian distances with the centroid and median methods in that the dendrograms, with unmodified data, are step-wise and very difficult to interpret. This was not observed with modified data, using euclidian distances, but was present in a few cases with the centroid method using unmodified data and the proper squared euclidian distances. Dendrograms produced using the *cosine* θ measure and Ward's method give results where the majority of the clusters combine at the lowest level of similarity, and hence are meaningless. Nonetheless, it takes little effort to try all distance or similarity measurements on the algorithms to see what happens, much as Jones (1988, p.16) suggests, even though the mathematics may not be rigorously adhered to, and some of the resultant dendrograms may be of limited use.

A third measure, allied to distance measurements, is a similarity measure known as *cosine* θ (or as *cos* η, or the *cosine* measure, either considered a shape measure or a pattern similarity measure: note 2, appendix). Imbrie and Purdy (1962) have used it in their study of bahamian carbonates and faunas. Zhang and Hofmann (1982) employed it in their study of lamina shape of Precambrian stromatolites. Ward (1985) has used the *cosine* θ measure to compare disjunct variables in Cretaceous communities in Canada; other references are given by Romesburg (1984, p. 109). An *a priori* assumption in this study was that this measure could possibly help in grouping similar faunas differing only in diversity, but this was not borne out.

Coefficients of association. Jaccard's coefficient of association was used as representative of the numerous coefficients of association that have been suggested. It may be noted here that it can only be used on absence–presence data, and it does not take into account the absence of a taxon in both collections being compared. Furthermore, this coefficient has long been used in ecology (Sneath and Sokal 1973, p.131), is assuredly one of the best known, and most widely used of its class (Lespérance and Sheehan 1988), although its shortcomings have previously been pointed out (Raup and Crick 1979). Archer and Maples (1987) have also questioned the utility of Jaccard's coefficient, based on a probabilistic (gaussian) approach (Anderberg 1973, p. 91 discusses only briefly this aspect). In any event, the use of Jaccard's coefficient is meant as a test of raw versus modified data.

Input order and clustering

Even though some methods of clustering require specific measurements of distance, all data sets 48, 46 and 40 were sequentially submitted to the seven clustering algorithms, with all three distance-similarity measures taken in turn. Both the raw and modified data were submitted to the same cluster techniques, and dendrograms of the results generated.

In order to judge if indeed clustering is chaotic, subjective criteria had to be devised. An obvious result of the great majority of the dendrograms generated was that the nine localities of the *Ml* Community (Watkins 1979, table 18), as well as the nine localities of the *Go* Community (Watkins 1979, table 17), with commonly the addition of the *Go1* (*Glassia obovata* fauna in the laminated shale facies, table 15), were correctly clustered together, at various levels of similarity depending on the measure used. The level of similarity defining each individual community was found, and the

TABLE 2. Comparison of clusters produced upon varying order of data entry on unmodified and modified absence–presence data

	ds	WPGMA U/M	UPGMA U/M	Single U/M	Complete U/M	Centroid U/M	Median U/M	Ward U/M
EUCLID	48	–/I	–/–	I/I	–/I	–/I	–/I	–/–
	46	–/I	–/–	I/I	–/I	I/I	I/I	–/–
	40	I/I	I/I	I/I	I/I	–/I	I/–	I/I
SEUCLID	48	I/I	–/–	I/I	–/I	–/I	I/I	–/–
	46	I/–	–/–	I/I	–/I	–/I	I/I	–/I
	40	–/I	I/I	I/I	I/I	I/I	–/I	I/I
COS *θ*	48	I/I	I/I	I/I	I/I	–/I	I/–	NA/NA
	46	–/I	I/I	I/I	I/I	–/I	–/I	NA/NA
	40	I/I	I/I	I/I	I/I	I/I	I/I	NA/NA

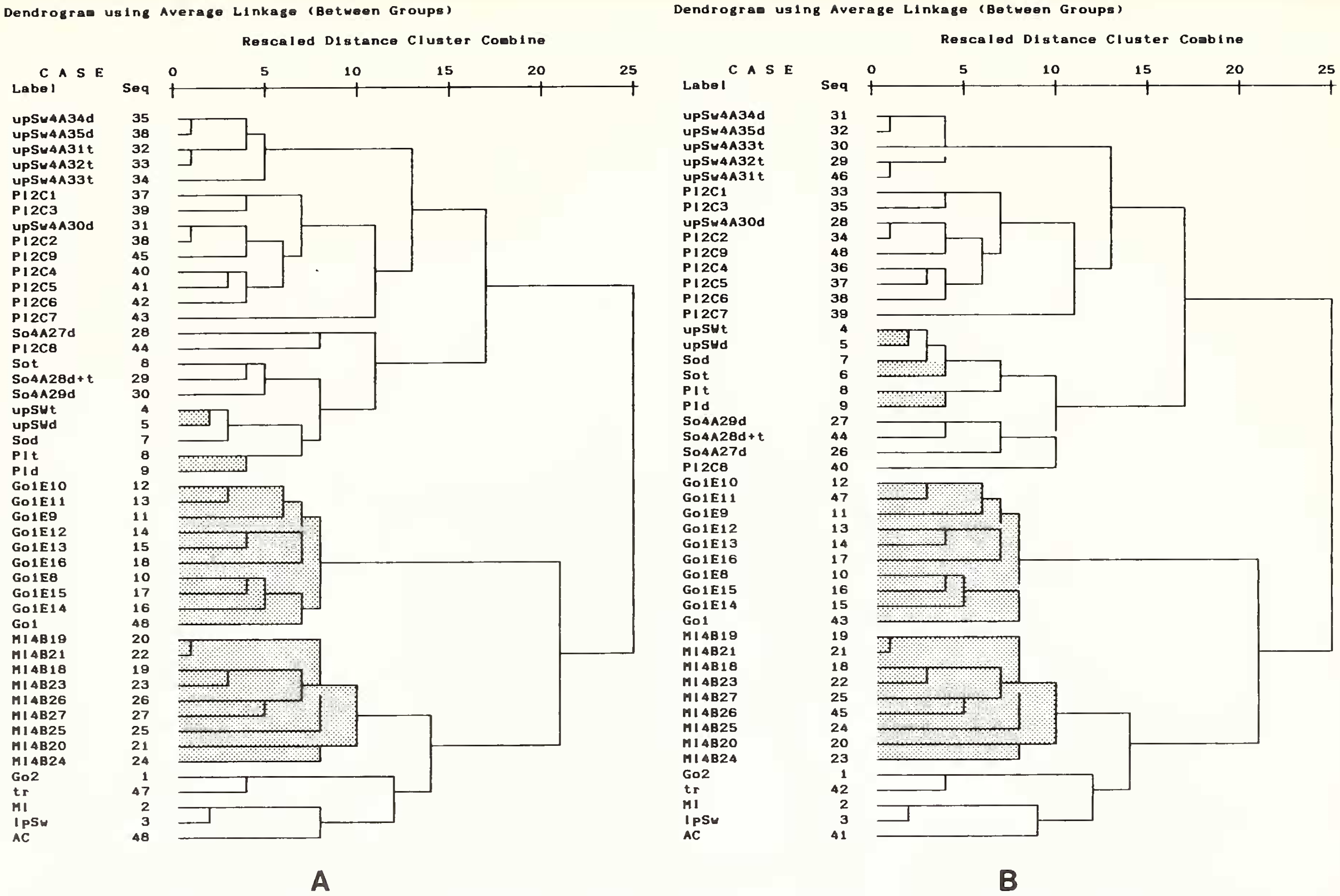
I: clusters identical; – clusters do not contain the same collections or faunas; NA: not applicable (clusters meaningless); ds: data set; U: unmodified data; M: data modified with Jaccard's coefficient; EUCLID: euclidian distance measurements; SEUCLID: squared euclidian distance measurements; COS *θ*: *cosine θ* similarity measure. Successive data sets derived from the preceding by the deletion of cumulative faunas (*Go2* and *Ml* from data set 48 to obtain data set 46; and deletion of disturbed and transported assemblages *upSw*, *So*, and *Pl* from data set 46 to obtain data set 40).

average between the two noted. This average was projected throughout the dendrogram (i.e. a phenon line was plotted) and used to assess if clusters of the data using different orders of entry were identical. All major clusters with identical or slightly higher levels of similarity than this phenon line were compared. Those that contained the same faunas (or collections) were judged to be clustered identically. In cases where the dendrograms obtained were step-wise (chaining of Jones 1988, p. 16), judgement was less structured but followed the same general principles.

Table 2 presents the results obtained upon varying the order of data entry on the raw and modified data. Obviously, chaotic clustering behaviour occurs, and is a significant characteristic to be reckoned with (text-fig. 1 is typical of the results obtained). This chaotic behaviour was not detected when using the single linkage method, but occurs with all other methods. It occurs equally abundantly whether euclidian or squared euclidian distance measurements are used, but less so when the *cosine θ* similarity measure is employed. Furthermore, the modified data clearly are less prone to chaotic behaviour than the raw data. Viewed somewhat differently, the centroid, median and Ward's methods perform no better with their mandatory squared euclidian distances.

A coefficient of association is thus useful, if only to decrease chaotic behaviour, contradicting one of the first assumptions of this study. The *cosine θ* similarity measure also appears almost ideal when compared to either of the other two distance measurements, particularly if the centroid, median and Ward's methods are restricted to their proper squared euclidian distance measurements. This improvement in non-chaotic behaviour is interpreted to be the result of the lesser, finite raw data consisting of zeros and ones, as compared with the more infinite numbers generated using a coefficient of association. Ten decimals were used to compute Jaccard's coefficient, thus greatly decreasing the chance of two points in hyperspace having the same coordinates (even though only a single quadrant, of 0 to 90°, is considered with the *cosine θ* measure), thus decreasing chaotic behaviour as described by Bayer (1985).

The most noteworthy results of table 2 are (a) the poor performance of the UPGMA method with euclidian and squared euclidian distances and the equally poor performance of Ward's method with all three distance-similarity measures, (b) the excellent performance of the single linkage method with all three distance-similarity measures, and (c) the equally excellent performance of the *cosine θ* similarity measure with unmodified and modified data with Jaccard's coefficient.



TEXT-FIG. 1. Unretouched computer printouts of dendrograms demonstrating chaotic clustering behaviour. Printouts obtained with a 9-pin dot-matrix printer employing emphasized combined with overstrike printing (a Citizen 120-D printer was used). Missing ascenders and descenders at the junction of some vertical and horizontal lines presumably due to the failure of the statistical program to use ASCII codes 169 and 170. 'Case labels' are identical to Watkins's (1979) association or sample numbers in his tables 15–20. 'Case seq' gives the input sequence of the data. Shading added to clusters reproducing Watkins's (1979) communities or his three cumulative disturbed-transported assemblages. *A*, using data sct 48A, modified with Jaccard's coefficient of association, euclidian distances and the UPGMA algorithm. *B*, as *A*, but using data set 48B.

CLUSTERING EFFICIENCY

Again, subjective criteria of clustering efficiency (in relation to the previously described communities) must be devised to judge cluster analyses against Watkins's (1979) communities. The make-up of the various data sets, as well as an overview of the communities, have been previously given, but additional considerations are essential to understand the criteria used.

Data set 48 presents an unrealistic situation of cluster analysis in which cumulative faunas and, partly, their constituent parts are considered in a single data set. Nevertheless, it does offer extremes in diversity and an opportunity of testing chaotic behaviour in cluster analysis. Efficiency of clustering was extended to this data set for the sake of completeness.

Data set 46 contains the results of a major conclusion submitted by Watkins (1979). He has submitted that transported assemblages are similar in content of epifaunal species to adjacent disturbed neighbourhood assemblages, that the basic community integrity of the transported assemblages is maintained (Watkins 1979, pp. 207–208), and that there was no significant difference between the two.

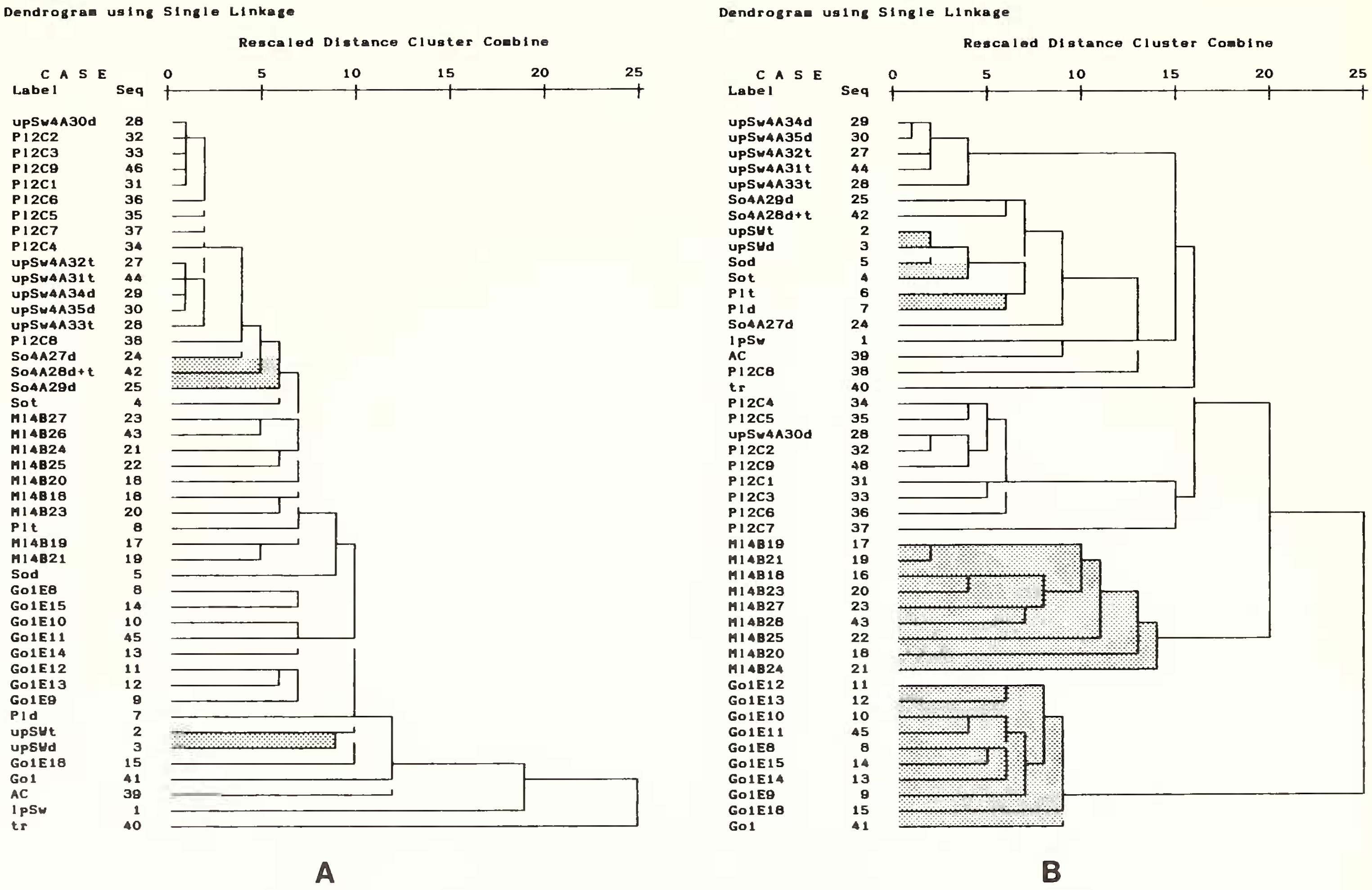
Data set 46 is more realistic than 48, but nonetheless suffers from the improbability of including cumulative taphonomic assemblages with individual collections and, partly, their cumulative faunas. Even so, the taphonomically separated cumulative faunas (Watkins 1979, table 16) do give more comprehensive cumulative faunas than the sum of the individual collections cited in succeeding tables. Data set 40 excludes these taphonomically separated faunas and is, with data set 46, not ideal as some sort of standard. Data sets 46 and 40 must consequently be used, at least as data for tests of clustering efficiency with sets of differing diversity.

TABLE 3. Clustering efficiency as judged from comparison with previously described communities or faunas of Watkins (1979)

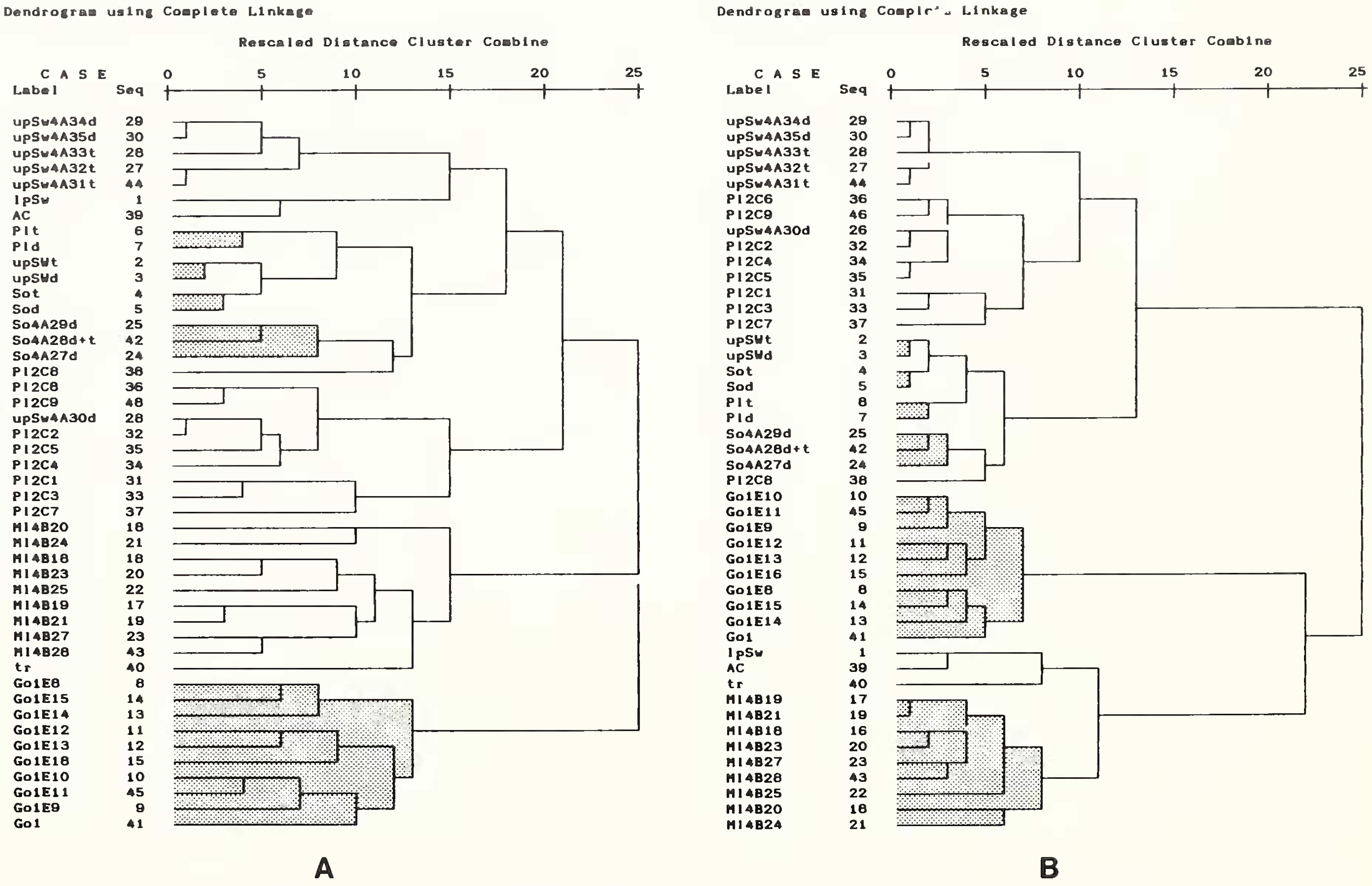
		WPGMA	UPGMA	Single	Complete	Centroid	Median	Ward
	ds	U/M	U/M	U/M	U/M	U/M	U/M	U/M
EUCLID	48A	NC/8	NC/NC	1/6	NC/8	NC/6	NC/NC	NC/NC
	46B	NC/8	NC/NC	1/6	NC/6	NC/6	NC/NC	NC/NC
	40A	NC/5	NC/NC	1/5	NC/5	NC/5	NC/NC	NC/NC
SEUCLID	48A	NC/NC	NC/NC	1/6	NC/8	NC/8	NC/8	NC/NC
	46B	NC/NC	NC/NC	2/7	NC/7	NC/7	NC/7	NC/NC
	40A	NC/NC	NC/NC	1/5	NC/5	NC/5	NC/5	NC/NC
COS θ	48A	NC/8	7/6	4/6	7/9	NC/7	NC/NC	NA/NA
	46B	NC/7	6/8	6/7	8/8	NC/6	NC/NC	NA/NA
	40A	NC/5	4/5	4/5	4/5	NC/4	NC/NC	NA/NA

NA: not applicable (clusters meaningless); NC: not considered; ds: data set, maximum possible scores on 48A = 12, 46B = 11 and 40A = 8; U: score on unmodified data; M: score on data modified with Jaccard's coefficient; EUCLID: euclidian distance measurements; SEUCLID: squared euclidian distance measurements; COS θ:*cosine* θ similarity measure. See table 2 for explanation of the derivation of the data sets, and the text for explanation of A and B suffixes.

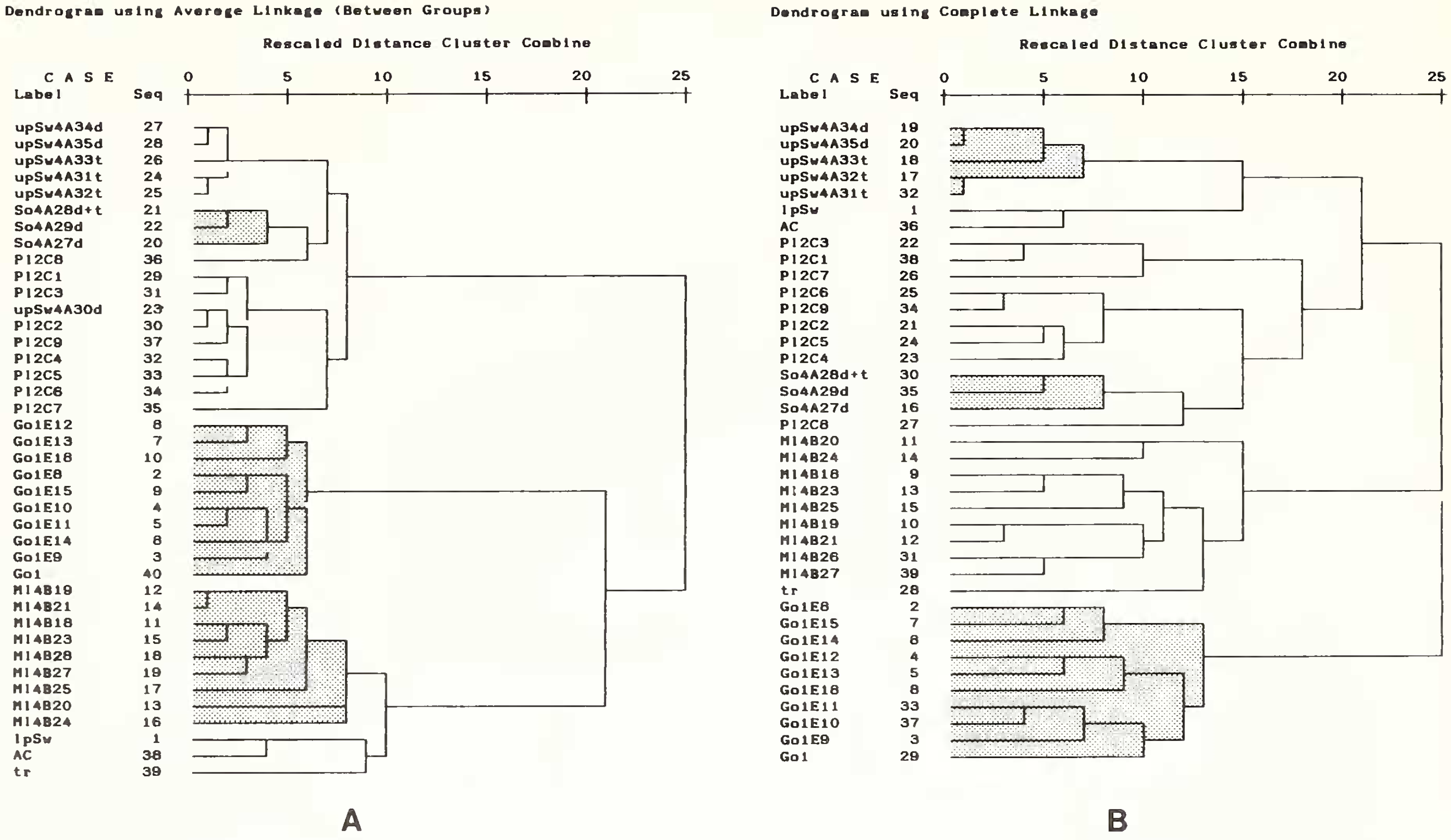
Keeping in mind the nature of the clustered data, and the fact that high diversity collections will cluster together before clustering with their constituent parts, or other low diversity collections (notes 3 and 4, appendix), it is possible to define *a priori* clustering results. A total of 12 sets of circumstances applicable to data set 48 can be envisioned (11 on data set 46 and 8 on data set 40); these are numbered (1) to (12) below. Clustering efficiency can be judged with the **requirement** that individual collections from specific localities assigned to specific communities (Watkins 1979, tables 17–20) be individually recognized; these are (1) the *Go* medium diversity community, (2) the *Ml* low



TEXT-FIG. 2. Unretouched computer printouts demonstrating increase in clustering efficiency with data modified with Jaccard's coefficient, otherwise as text-fig. 1. *A*, using unmodified absence–presence data, data set 46B, squared euclidian distances and the single linkage algorithm. *B*, same as *A*, but data set modified with Jaccard's coefficient.



TEXT-FIG. 3. Unretouched computer printouts demonstrating equally good clustering efficiency using the *cosine θ* similarity measure on unmodified and modified data, otherwise as text-fig. 1. *A*, using unmodified data set 46B and the complete linkage algorithm; position of transitional fauna (*tr*) and lower phase of *Sphaerirhynchia wilsoni* (*lpSw*) fauna are notable. *B*, as *A*, but with data set 46B modified with Jaccard's coefficient.



TEXT-FIG. 4. Unretouched computer printouts demonstrating typical clustering results and increasing efficiency in clustering, otherwise as text-fig. 1. *A*, typical results on data modified with Jaccard's coefficient, using data set 40A, the *cosine θ* measure and the UPGMA clustering algorithm. *B*, better clustering of Watkins's (1979) communities, using unmodified data set 39C, *cosine θ* and the complete linkage algorithm.

to medium diversity community, (3) the *So* medium diversity community, (4) the *upSw* low diversity community and (5) the *Pl* low diversity community; **expectation** that the cumulative faunas of the disturbed neighbourhood assemblages and transported assemblages of the same communities be clustered together at very high levels of similarity, and ideally, should form binary clusters; such should be (6) *upSw*t-*upSw*d, (7) *So*t-*So*d and (8) *Pl*t-*Pl*d; and **hope** (because of the significant differences in diversity) that the four medium to high diversity cumulative faunas retained from table 15 of Watkins (1979) cluster with their respective counterparts given by specific collections and localities; such should be (9) *Go*1 with the *Go* community, (10) *lpSw* with *upSw*, (11) *Ml* with *Ml* individual collections and (12) the *tr* fauna with either the *Go* of *Ml* communities or, ideally, in an intermediate position. The *AC* community has been considered anomalous (Watkins 1979, p. 231) and should not, *a priori*, cluster with any particular community.

The same dendrograms generated for testing chaotic behaviour were used to test these assumptions of efficiency. Because of the chaotic behaviour of some algorithms with specific distance-similarity measures, only those cases exhibiting no chaotic behaviour were considered (e.g. WPGMA with euclidian distance measurements was considered only with modified data).

Table 3 presents the scores obtained from the previous 8 to 12 assumptions, depending on the data set. These results are subjective judgments, and should be viewed as such. Interpretation of table 3 must consider which type of data is used for clustering. In the case of unmodified absence–presence data, clustering efficiency is very poor if euclidian or squared euclidian distance measurements are used and, in fact, best not employed. *Cosine θ* similarity measurements give greatly improved results in efficiency (text-fig. 2).

The choice of measures and methods increases greatly with data modified with Jaccard's coefficient, and all are approximately equally efficient when chaotic behaviour does not occur. A most surprising conclusion, however, resides in the fact that *cosine θ* measures with raw or modified data give nearly equally good results (text-fig. 3). Routinely applicable combinations are *cosine θ* with UPGMA, single and complete methods, with raw or modified data. Such a combination with unmodified data is particularly well suited to obtain quick results and is recommended as ideally suited, at least, for first approximations in community analyses (text-fig. 4A is typical).

Nonetheless, even the highest scores of table 3 do not reach the maximum permissible ones. A serious deficiency of the cluster analysis is the failure to recognize, in all the dendrograms generated, the low diversity *upSw* and *Pl* communities. The *upSw*4A30d collection always clustered with the *Pl* series. The fauna of the *upSw*4A30d locality has $d = 6$, the lowest of the *upSw* localities, and five of its constituent taxa are present in an average of 74% of the *Pl* collections. It is consequently far from surprising that this locality clusters with the *Pl* series; it is logically part of the *Pl* community, even though it lies immediately below the *So* community. It was a simple matter to exclude the *upSw*4A30d locality from sets 40A and 40C, to form data set 39 (39A and 39C), which was submitted to cluster analysis using the single linkage algorithm with euclidian, squared euclidian and the *cosine θ* measure; this last measure was also used with the UPGMA and complete linkage algorithms. Chaotic behaviour was not detected, and clustering was most efficient, again, with the *cosine θ* measure; complete linkage yielded the best results (text-fig. 4B). Text-figures 3A and 4B both contain the correct positioning of the *lpSw* fauna, near *upSw*, and the transitional fauna (*tr*) occurs close to the *Ml* fauna (by definition, the *tr* fauna is intermediate between the *Ml* and *Go* faunas), and, in fact, it destroys the *Ml* cluster.

In a study which pretends to be objective, the procedure outlined above to obtain data set 39 is not without serious reservations (i.e. the procedure is subjective); additional speculation appears unwarranted (but see note 5 of the appendix).

DISCUSSION

The pragmatic approach utilized here to test chaotic behaviour and clustering efficiency of previously described communities from part of the Welsh Borderland is probably the most realistic method of unravelling the intricacies and uncertainties associated with cluster analysis as presently understood. Some specific methods are judged to be safe and do show positive results, *with the specific data set employed.* Additional testing is necessary to extend their applicability. It would appear, for instance, that unmodified presence–absence data, combined with the median method and squared euclidan distances employed by Lespérance and Sheehan (1988, appendix 2) to recognize previously defined communities by them is a spurious result, in view of tables 2 and 3. Nonetheless, chaotic behaviour occurs in cluster analysis, even if some specific algorithms and distance–similarity measurements appear preferable to others. Future investigators should therefore check the trustworthiness of their dendrograms by varying their order of data entry to test if chaotic behaviour occurs with their specific data.

Lespérance and Sheehan (1988) have contrasted the methodologies employed during 'classical' community definition, non-parametric cluster analysis and parametric statistical methods (as discriminant analysis); this is pertinent to this contribution, but need not be repeated here. Results obtained in the present study could be interpreted to mean that the 'classical' community approach is vindicated by cluster analysis. On the other hand, the same results could also indicate that cluster analysis is more rigorous, gives reproducible results, and is as precise as the 'classical' approach.

The data set employed here is typical of 'middle' Palaeozoic data, and of such magnitude as to be representative of situations when cluster analysis is envisioned, i.e. when visual examination does not reveal the underlying structure of the data. Communities are recurring associations of taxa and cluster analysis is a statistical method specifically construed to group samples, and hence recognize communities. It is consequently worthy of more extended usage. Presently available PCs and software greatly facilitate this.

Acknowledgements. Peter M. Sheehan, of the Milwaukee Public Museum, convinced the writer a number of years ago of the necessity of cluster analysis in specific circumstances. R. Watkins (of the same museum) and A. J. Boucot (Oregon State University) commented on a nearly complete draft of this study. Operating grants from the Natural Sciences and Engineering Research Council of Canada were essential to this study, and are gratefully acknowledged.

REFERENCES

ANDERBERG, M. R. 1973. *Cluster analysis for applications.* 359 pp. Academic Press.

ANTIA, D. D. J. 1980. Shell laminae and shell orientation in the Upper Silurian, Overton Formation, U.K. *Palaeogeography, Palaeoclimatology, Palaeoecology*, **32,** 119–133.

ARCHER, A. W. and MAPLES, C. R. 1987. Monte Carlo simulation of selected binomial similarity coefficients (I): effect of number of variables. *Palaios*, **2,** 609–617.

BAARLI, B. G. 1987. Benthic faunal associations in the Lower Silurian Solvik Formation of the Oslo-Asker Districts, Norway. *Lethaia*, **20,** 75–90.

BAYER, U. 1985. *Lecture Notes in Earth Sciences. 2 Pattern recognition problems in geology and paleontology.* FRIEDMAN, G. M. and SEILACHER, A. (eds.), vii + 229 pp. Springer-Verlag.

BOYCE, A. J. 1969. Mapping diversity: a comparative study of some numerical methods, pp. 1–31. *In* COLE, A. J. (ed.). *Numerical taxonomy*, Academic Press, New York.

CALEF, C. E. and HANCOCK, N. J. 1974. Wenlock and Ludlow marine communities in Wales and the Welsh Borderland. *Palaeontology*, **17,** 779–810.

CHEETHAM, A. H. and HAZEL, J. E. 1969. Binary (presence–absence) similarity coefficients. *Journal of Paleontology*, **43,** 1130–1136.

CHERNS, L. 1988. Faunal and facies dynamics in the Upper Silurian of the Anglo-Welsh Basin. *Palaeontology*, **31,** 451–502.

EVERITT, B. 1980. *Cluster analysis.* Second edition. Heinemann Educational Books, London (Halsted Press, J. Wiley and Sons), 136 pp.

HEWITT, R. A. and WATKINS R. 1980. Cephalopod ecology across a late Silurian shelf tract. *Neues Jahrbuch für Geologie und Paläontologie, Abhandlungen*, **160,** 96–117.

HOLLAND, C. H. 1980. Silurian series and stages: decisions concerning chronostratigraphy. *Lethaia*, **13,** 238.

——, LAWSON, J D. and WALMSLEY, V. G. 1963. The Silurian rocks of the Ludlow district. *Bulletin of the British Museum (Natural History), Geology*, **8,** 95–171.

HOLLAND, C. H., LAWSON, J D., WALMSLEY, V. G. and WHITE, D. E. 1980. Ludlow stages. *Lethaia*, **13,** 268.

IMBRIE, J. and PURDY, E. G. 1962. Classification of modern bahamian carbonate sediments, pp. 253–272. *In* HAM, W. E. (ed.). *Classification of carbonate rocks. American Association of Petroleum Geologists, Memoir*, **1,** 279 pp.

JONES, B. 1988. Biostatistics in paleontology. *Geoscience Canada*, **15(1),** 3–22.

LAWSON, J. D. 1975. Ludlow benthonic assemblages. *Palaeontology*, **18,** 509–525.

LEGENDRE, L. and LEGENDRE, P. 1983. *Developments in environmental modelling, 3 Numerical ecology.* Elsevier, Amsterdam, xvi + 419 pp.

LEGENDRE, P., DALLOT, S. and LEGENDRE, L. 1985. Succession of species within a community: chronological clustering, with applications to marine and freshwater zooplankton. *The American Naturalist*, **125,** 257–288.

LESPÉRANCE, P. J. and SHEEHAN, P. M. 1988. Faunal assemblages of the Upper Gaspé Limestones, Early Devonian of Eastern Gaspé, Quebec. *Canadian Journal of Earth Sciences*, **25,** 1432–1449.

MIKULIC, D. G. and WATKINS, R. 1981. Trilobite ecology in the Ludlow Series of the Welsh Borderland, pp.101–117. *In* GRAY, A. J., BOUCOT, A. J. and BERRY, W. B. N. (eds.). *Communities of the past.* Hutchinson Ross, Stroudsburg, Pennsylvania.

MILLIGAN, G. W. 1980. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, **45,** 325–342.

—— and ISAAC, P. D. 1980. The validation of four ultrametric clustering algorithms. *Pattern Recognition*, **12,** 41–50.

NORUSIS, M. J. 1988. *SPSS/PC+ Advanced Statistics V2.0.* SPSS Inc., Chicago. [Pages A1–A2, B1–B214, C1–C91, D1–D30, E1–E6.]

RAUP, D. M. and CRICK, R. E. 1979. Measurement of faunal similarity in paleontology. *Journal of Paleontology*, **53**, 205–212.

ROMESBURG, H. C. 1984. *Cluster analysis for researchers.* Lifetime Learning Publications, Belmont, California, 334 pp.

SNEATH, P. H. A. and SOKAL, R. R. 1973. *Numerical taxonomy – The principles and practice of numerical taxonomy.* W. H. Freeman and Co, San Francisco, xv + 573 pp.

WARD, J. E. JR. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, **58,** 236–244.

WARD, P. D. 1985. Upper Cretaceous (Santonian-Maastrichtian) molluscan faunal associations, British Columbia, pp. 397–420. *In* BAYER, U. and SEILACHER, A. (eds.). *Lecture Notes in Earth Sciences. 1 Sedimentary and Evolutionary Cycles,* Springer-Verlag.

WATKINS, R. 1978. Bivalve ecology in a Silurian shelf environment. *Lethaia*, **11,** 41–56.

—— 1979. Benthic community organization in the Ludlow Series of the Welsh Borderland. *Bulletin of the British Museum* (*Natural History*), *Geology*, **31,** 175–280.

—— and AITHIE, C. J. 1980. Carbonate shelf environments and faunal communities in the Upper Bringewood Beds of the British Silurian. *Palaeogeography*, *Palaeoclimatology*, *Palaeoecology*, **29,** 341–368.

ZHANG, Y. and HOFMANN, H. J. 1982. Precambrian stromatolites: image analysis of lamina shape. *The Journal of Geology*, **90**, 253–268.

PIERRE J. LESPÉRANCE

Département de géologie
Université de Montréal
Casier Postal 6128
Montréal H3C 3J7 – Canada

Typescript received 7 November 1988
Revised typescript received 17 April 1989

APPENDIX – NOTES

1. *Jaccard's coefficient.* This is defined as the number of taxa in common between any two collection *p* and *q*, divided by the number of taxa not shared by the two collections. This can be expressed as

$$\frac{c}{a+b},$$

where *c* is the number of taxa in common between collections *p* and *q*, *a* the number of taxa restricted to collection *p*, and *b* the number of taxa restricted to collection *q*. As the total number of taxa is more readily available from the raw data than taxa restricted to either collection, it is more convenient to calculate Jaccard's coefficient as

$$\frac{c}{d+e-c},$$

where d is the total number of taxa in collection p, and e the total number of taxa in collection q.

As each and every collection must be individually compared, Jaccard's coefficient consequently requires, where n = number of collections,

$$(n-1)(n/2)$$

times the calculation of the formula (Imbric and Purdy 1962, p. 258), that is, number of rows within a spreadsheet. As a first step in calculating the coefficient, the number of taxa common between each and every collection was isolated. With the spreadsheet used, the following formula (in 'Symphonese') was employed, assuming that each collection (1, 2, ... n) is in a row, that collection p is in a row 2, q is in row 3, that labels of the taxa within the data set to be analysed are in cells A1, B1, ..., and absence–presence data (0 or 1) are in rows 2, 3, ... $n+1$:

$$@\mathrm{IF}((\mathrm{A}2+\mathrm{A}3)=2,1,0)+@\mathrm{IF}((\mathrm{B}2+\mathrm{B}3)=2,1,0)+\ldots,$$

which reads: if the sum of cells A2 and A3 is 2, return 1, otherwise return 0, plus .... This is the most time-consuming process of calculating the coefficient and, furthermore, is the one which consumes the greatest amount of memory. To calculate the 48 collections and 112 taxa of data set 48A or 48B within a single file would require, approximately, 3·5 Mb of mainboard and expanded memory; it was easier to split the calculations into three distinct files of almost 1·4 Mb each. Jaccard's coefficient was finalized in another file, after extracting and combining the values obtained from the previous three files.

2. *The cos θ (cosine theta) similarity measure.* The origin of this angular measure of similarity is uncertain. Boyce (1969, p. 3) traces it back to 1946 and attributes it to A. Bhattacharyya. Imbrie and Purdy (1962, p. 257) introduced vector notation in their study and used the following angular measure of similarity:

$$\frac{\sum_{i=1}^{n} x_{ip} . x_{iq}}{\sqrt{\sum_{i=1}^{n} x_{ip}^{2} \sum_{i=1}^{n} x_{iq}^{2}}},$$

where n are rock properties (here considered as absence–presence data), p and q are any pair of samples, and a generalized observation x_{ij} on the i'th property of the j'th sample ($i = 1, 2, \ldots n; j = 1, 2, \ldots N$ samples). This measure of similarity replaces euclidian, or squared euclidian distances (or other distance or similarity measures) in cluster analysis. The *cos θ* measure is not discussed in some standard texts. Anderberg (1973, p. 71) presents perhaps the most extensive discussion. The SPSS/PC+ statistical package offers it as a choice within the clustering module.

Boyce (1969) uses the formula above to define his *cosine of angle*, considers it a measure of resemblance and, more specifically, a measure of similarity in shape. Sneath and Sokal (1973, p. 172) define *cos η* with the same formula quoted above, and follow Boyce (1969). Anderberg (1973) uses the same formula, names the result the cosine of the angle between the vectors, and views it as a measure of similarity between two vectors. The SPSS/PC+ Advanced Statistics manual (Norusis 1988, p. B-86) again uses the same formula, names the cosine measure the *cosine of vectors of variables*, and considers it a pattern similarity measure.

3. *Clustering truism 1.* Diversity directly affects clustering, e.g. low diversity collections will cluster together, while high diversity ones will do likewise, before clustering with the low diversity ones.

Consider the worst case situation in which four collections $a, \ldots d$ with, respectively, diversities of 8, 10, 44 and 55, where the ratios of $a{:}b$ are as $c{:}d$. Furthermore, assume that any smaller collection is totally included in any larger collections. These data, modified with Jaccard's coefficient of association produce the following symmetrical matrix:

$$\begin{array}{lllll}
 & a & b & c & d \\
a & 1{\cdot}0 & & & \\
b & 0{\cdot}800000 & 1{\cdot}0 & & \\
c & 0{\cdot}181818 & 0{\cdot}227273 & 1{\cdot}0 & \\
d & 0{\cdot}145455 & 0{\cdot}181818 & 0{\cdot}800000 & 1{\cdot}0
\end{array}$$

It is obvious from above that collections $a-b$ and $c-d$ will cluster together before any other collection.

4. *Clustering truism 2.* Parts of a whole will cluster together before they cluster with the whole, e.g. individual collections of a community will cluster together before clustering with the cumulative fauna of the community.

This is but a special case of the preceding. Consider that collections *a*, *b* are collections within the cumulative fauna *c*, and that fauna *d* is a cumulative fauna from another community. Collections *a*, *b* are obviously contained in *c*, but if fauna *d* is from another community, it should have less than 50% taxa in common with *c*, say in the worst case, 45% (= 25 taxa). Assuming that *a*, *b* are also contained within *d*, these data produce:

	a	*b*	*c*	*d*
a	1·0			
b	0·80000	1·0		
c	0·181818	0·227273	1·0	
d	0·145455	0·181818	0·337838	1·0

It is clear that collections *a*—*b* will cluster first, followed by *c*—*d* at lower levels of similarity. In fact, with the above matrix, a dendrogram of the clustering (using all the measures and methods outlined in the main text) shows that the *c*—*d* cluster joining the *a*—*b* cluster only at the lowest level of similarity.

5. *Use of the Manhattan (city-block) distance measure.* Usc of the infrequently employed Manhattan, or city-block, distance measurement (instead of euclidian or squared euclidian distances; see Legendre and Legendre 1983, p. 198), in conjunction with unmodified data in sets 40A, 40C, 39A and 39C, with complete linkage and Ward's algorithms, gave increasingly better clustering. Perfect clustering was obtained when localities *Pl*2C7 and *Pl*2C8 were deleted from data set 39, again using complete linkage and Ward's methods with unmodified data. Chaotic behaviour did not occur with any of the three data sets, in the specific circumstances described above, but occurred with unmodified data sets 48 and 46.